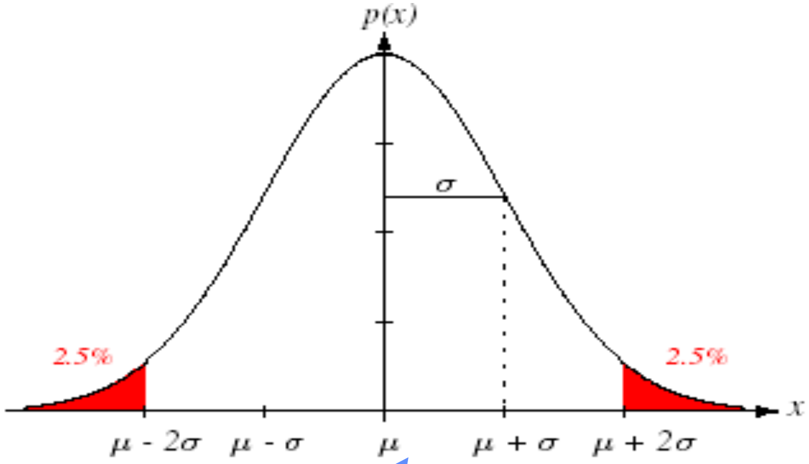
Part 4

Elements of parametric pattern classifiers: the Gaussian classifier

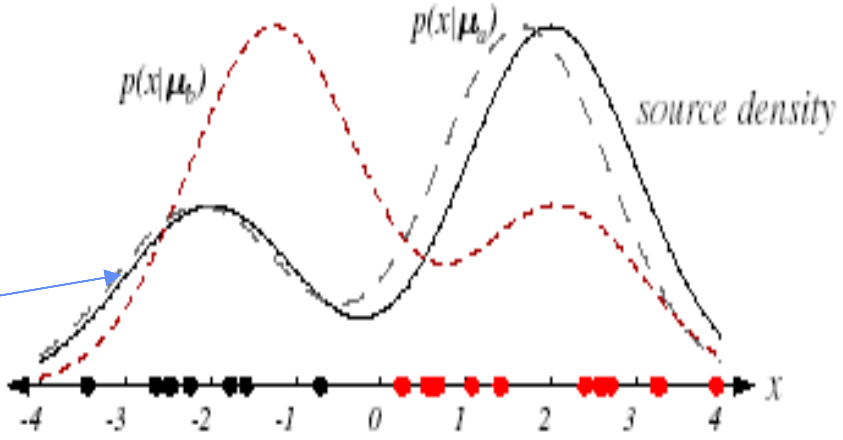
Parametric models

- Parametric methods assume that the class-conditional density function $p(\mathbf{x}/\boldsymbol{\omega}_i)$ has a known parametric form with some unknown parameters.
- Example: mono-dimensional Gaussian model: $p(x/\boldsymbol{\omega}_i)= p(x)=N(\mu, \sigma)$

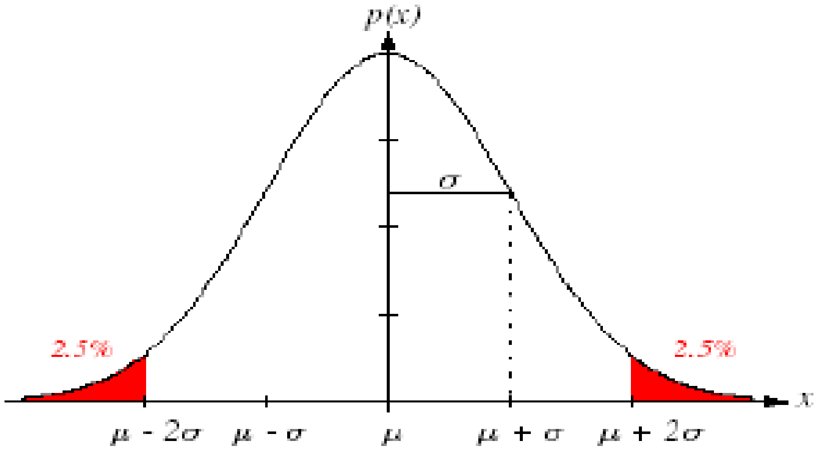


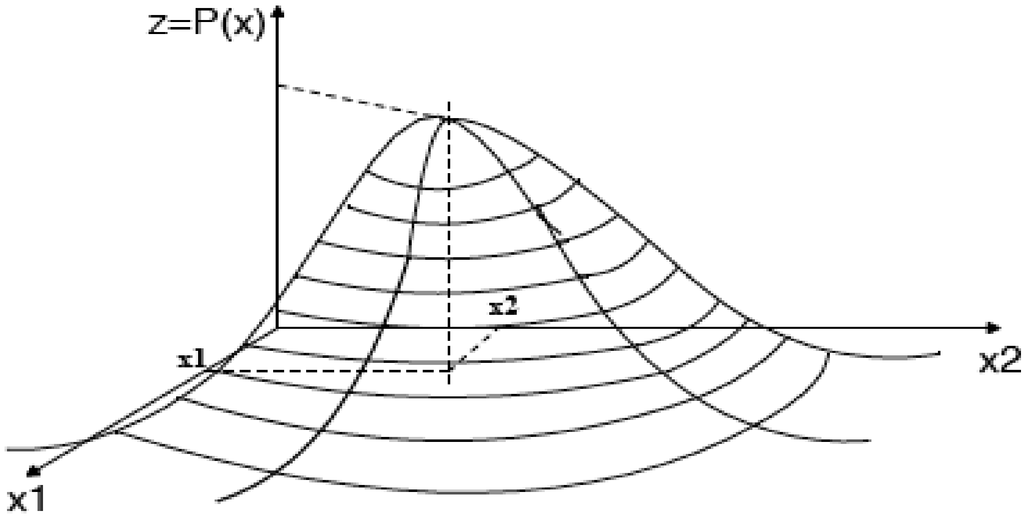
If the model of $p(x/\boldsymbol{\omega}_i)$ is Gaussian (see on the left), the problem is only to estimate the two parameters μ and σ.

- **Two main kinds of parametric models:**
  - Single "component" (also called mono-modal)
  - Mixture of components (also called multi-modals)



An example of single-component model: the Gaussian model





Why is the Gaussian model so widely used?

✓Several natural and/or artificial phenomena fit the Gaussian model (even scores of a student examination are roughly "normally-distributed"…)

✓Central Limit Theorem: the sum of N independent random variables will lead to a Gaussian distribution for $N \rightarrow +\infty$

✓In several pattern recognition tasks, a pattern can be regarded as an ideal 'prototype' corrupted by a sum of random and independent noisy sources

✓In some cases the distribution is not Gaussian, but it can be approximated by a sum of Gaussian functions.

Parametric form of the Gaussian model

- One-dimensional case:

$$p(x) = N(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\mu \equiv \varepsilon[x] = \int_{-\infty}^{+\infty} xp(x)dx$$

$$\sigma^2 \equiv \varepsilon\left[(x-\mu)^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx$$

- Multidimensional case:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)\right]$$

x and μ are column vectors of *d* components, Σ is the **covariance matrix** *d* x *d*, |Σ| and Σ^{-1} are its determinant and its inverse.

$$\mu \equiv \varepsilon[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})d\mathbf{x}$$

$$\Sigma \equiv \varepsilon\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^t\right] = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t p(\mathbf{x})d\mathbf{x}$$

$$\sigma_{ij} \equiv \varepsilon\left[(x_i-\mu_i)(x_j-\mu_j)\right]$$

Remarks on the covariance matrix

Properties of Σ:

- Σ is a symmetric matrix: $\Sigma = \Sigma^t$;
- Σ is a semi-positive definite matrix. However, in order to obtain a well-defined Gaussian probability density function, Σ must be positive definite (in fact, the expression of $p(\mathbf{x})$ involves the inverse of Σ and the division by the determinant $|\Sigma|$).

Independent random variables::

- Given the covariance matrix Σ

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2d} \\ \ldots & \ldots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$

- If $\sigma_{ij} = 0$, then the random variables x_i and x_j are uncorrelated and, being Gaussian distributed, they are also independent;
- If $\sigma_{ij} = 0$ for all $i \neq j$ (that is, if Σ is a diagonal matrix), we have:

$$p(\mathbf{x}) = p(x_1)\, p(x_2)\, ...p(x_d)$$

Gaussian class-conditional probability density

So far we have considered a generic Gaussian feature vector $\mathbf{x}$. Now consider a multi-class setting, in which a feature vector $\mathbf{x}$ is Gaussian and it is conditioned to a class ω_i, $i = 1, 2, \ldots, c$:

- $p(\mathbf{x} | \omega_i) = N(\mathbf{m}_i, \Sigma_i)$, $i = 1, 2, \ldots, c$;
- more explicitly:

$$p(\mathbf{x} \,|\, \omega_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^t \Sigma_i^{-1} (\mathbf{x} - \mathbf{m}_i) \right]$$

- $\mathbf{m}_i$ is the mean of $\mathbf{x}$ conditioned to ω_i: $\mathbf{m}_i = E\{\mathbf{x} | \omega_i\}$;
- Σ_i is the covariance matrix contitioned to ω_i:

$\Sigma_i = \mathrm{Cov}\{\mathbf{x} | \omega_i\} = E\{(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \,|\, \omega_i\}$

- by the theorem of total probability, the p.d.f. (probability density function) of x is a linear combination of Gaussian p.d.fs. (*Gaussian mixture*):

$$p(\mathbf{x}) = \sum_{i=1}^{c} P_i \, p(\mathbf{x} \,|\, \omega_i)$$

Gaussian classifiers

- Let us now consider which classifiers - and in particular which types of *discriminant functions* - may be obtained by assuming a Gaussian model of the data

- This analysis is important to understand the "complexity" of the classification tasks that can be solved by assuming a Gaussian model of the data

- We will also see the different types of classifiers that are obtained by different hypotheses (or "simplifications") on the parameters of the Gaussian model

If $p(\mathbf{x}/\omega_i) = N(\boldsymbol{\mu}_i, \Sigma_i)$

$$g_i(\mathbf{x}) = \ln\big(p(\mathbf{x}/\omega_i)\big) + \ln\big(P(\omega_i)\big)$$

The associate discriminant function is:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- The fundamental parameter that governs the "complexity" of the Gaussian model is obviously the matrix Σ; we will see different classifiers that are obtained under different assumptions about the structure of Σ.

- First, we see how one can estimate the mean value $\boldsymbol{\mu}_\mathbf{i}$ and the covariance matrix $\Sigma_\mathbf{i}$.

Quick notes on Maximum likelihood parameter estimation

- Given that samples are independent, we can write:

$$p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta) \ \text{likelihood of } \theta \text{ with respect to } D$$

- The **maximum likelihood estimate of θ is the value** $\theta = \hat{\theta}$ that maximizes p(*D*| θ). Intuitively, this estimate corresponds to the value θ so that p(**x**| θ) “fits” at the best patterns in ***D***.

Remember that $p(x/\omega_i)$ is the likelihood of x given the class ω_i.

➢*See the next slides for a qualitative explanation of the concept of “likelihood”*

- If p(*D*| θ) is a differentiable function, we can find the maximum likelihood parameters $(\theta_1,\ldots,\theta_p)$ by the standard methods of differential calculus.
- Let be

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^t$$

$$\text{We define the } l(\theta) = \log\left[p(D \mid \theta)\right] = \sum_{k=1}^{n} \log\left[p(\mathbf{x}_k \mid \theta)\right]$$

Logaritmic likelihood

Maximum likelihood parameter estimation

- The solution to the estimation problem provided by the equation

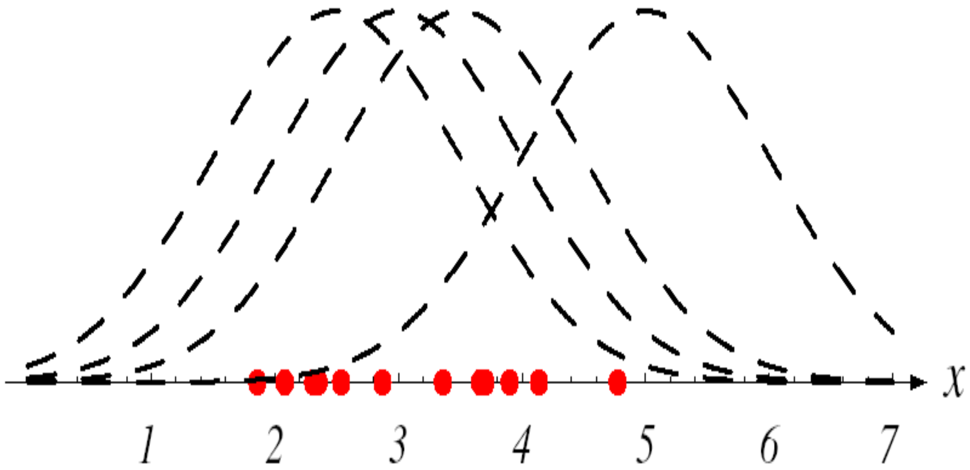
$$\nabla_\theta l(\theta) = 0$$

could represent a global maximum, a local maximum or minimum, or, more rarely, an inflection point .

We might have to check each solution individually, or calculate second derivatives. Clearly, it is necessary to look for the values of maximum even among the frontier points of the domain.

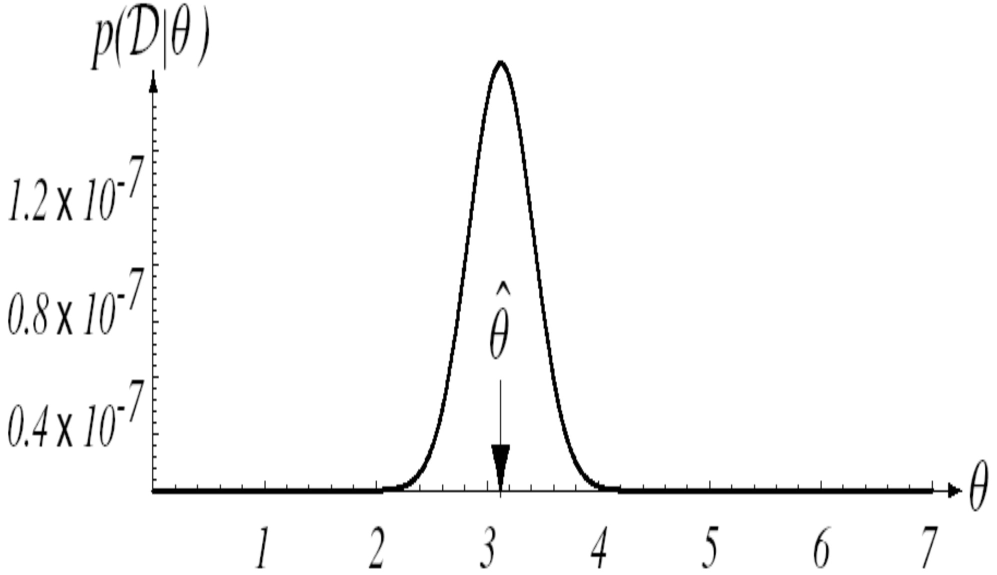
➢It should be noted that the solution found is only an *estimate* of θ. It is obvious because we are using a *particular limited set* of samples (***D***) to calculate it.

Qualitative example of the concept of likelihood



On the left (in red) we show a set *D* of patterns drawn from a Gaussian distribution with known variance but unknown mean value, which has to be estimated

What is the most likely mean value?



The figure on the left shows the likelihood function for the above example.

It is intuitive that θ =3 is the most likely mean value for the different Gaussian functions with equal variance.

Open issues:

-how to estimate the value?

-what is the “variance” of the estimate?

Maximum likelihood estimation of Gaussian parameters

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} \left(\mathbf{x}_k - \hat{\mu}\right)\left(\mathbf{x}_k - \hat{\mu}\right)^t$$

One dimensional feature space:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} \left(x_k - \hat{\mu}\right)^2$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n} \left(x_k - \hat{\mu}\right)^2$$ Unbiased estimator

Discriminant functions for Gaussian density

- Assume that we use the estimates we have seen for μ_i and Σ_i, and we estimate the prior probability $P(\omega_i)$ as the fraction of patterns of ***D*** that belongs to ω_i.

- Let us now consider which classifiers, and in particular which types of discriminant functions, may be obtained by assuming a Gaussian model of the data.

- This analysis is important to understand the "complexity" of the classification tasks that can be solved by assuming a Gaussian model of the data.

- We will also see the different types of classifiers that are obtained by different hypotheses (or "simplifications") on the parameters of the Gaussian model

If $p(\mathbf{x}/\boldsymbol{\omega}_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

The associate discriminant function is:

$$g_i(\mathbf{x}) = \ln\left(p(\mathbf{x}/\omega_i)\right) + \ln\left(P(\omega_i)\right)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\boldsymbol{\omega}_i)$$

$$P(\omega_i) = \frac{n_i}{\sum_j n_j}$$

- The fundamental parameter that governs the "complexity" of the Gaussian model is obviously Σ; in the next slides, we analyse the different classifiers that are obtained under different assumptions about the structure of Σ.

Gaussian model: case $\Sigma_i=\sigma^2 I$

- $\boldsymbol{\Sigma_i=\sigma^2 I}$ means that
  - The "feature" are statistically independent and have the same variance.
  - The data ("pattern") form hyper-spherical "clusters" (groups) of identical size, and centers μ_i
- In this simple case we obtain: $\left|\boldsymbol{\Sigma}_i\right| = \sigma^{2d} \quad \boldsymbol{\Sigma}_i^{-1} = \left(1/\sigma^2\right)\mathbf{I}$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

Then, the above $g_i(\boldsymbol{x})$ can be rewritten as:

$$g_i(\mathbf{x}) = -\frac{\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2}{2\sigma^2} + \ln\left(P(\omega_i)\right)$$

Gaussian model, with $\Sigma_i=\sigma^2 I$, the linear classifier

- Rewriting the $g_i(\mathbf{x})$ above and noting that the term ($\mathbf{x}^t\mathbf{x}$) is the same for all values of i, we obtain the *linear* discriminant function

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \qquad w_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i; \quad w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

w_{i0} is called *threshold* or "*bias*" for the i-th class

- The decision boundary is a portion of hyperplane of dimension d-1 defined by

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$ for classes with the highest posterior probability

In this case the equation of hyperplanes can be written as

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$$

where

$$\begin{cases} \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 = \dfrac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \dfrac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2}\ln\dfrac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{cases}$$

Gaussian model, $\Sigma_i = \sigma^2 I$, Linear Classifier

- The hyperplane that separates the regions R_i and R_j is orthogonal to the line joining the means

- The values of $P(\omega_i)$ and $P(\omega_j)$ determine the position of the point $\mathbf{x}_0$ in which the hyperplane passes

- Special case: $P(\omega_i) = P(\omega_j)$ for each class

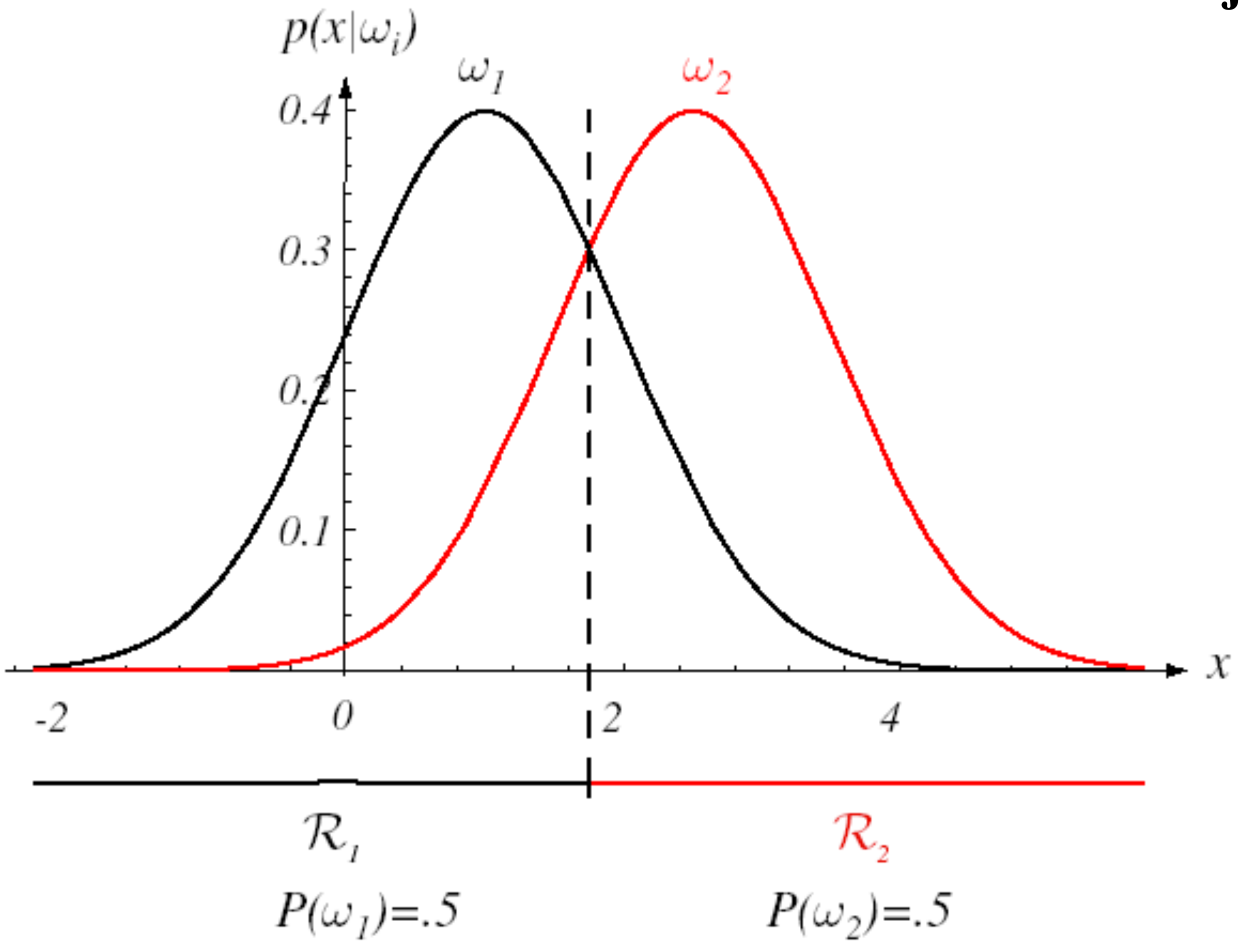
  - $P(\omega_i)$ becomes irrelevant to the classification

  - The discriminant function becomes

$$g_i(\mathbf{x}) = -\left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

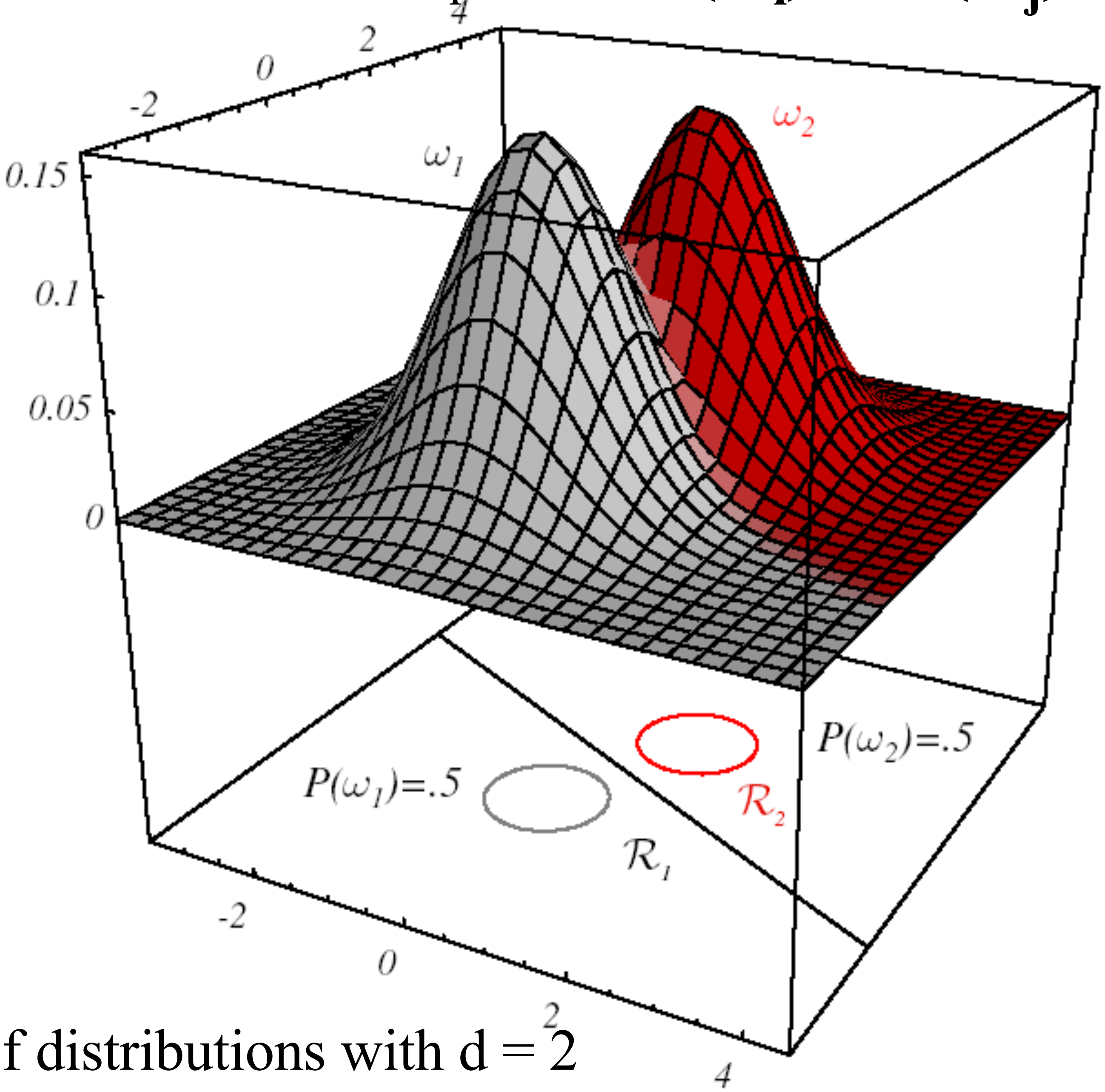
  - This classifier is called *minimum distance classifier* and it is used in the classification procedure called template matching (where each class is represented by its “prototype” μ_i)

Example for the case $\Sigma_i = \sigma^2 I$, $P(\omega_i) = P(\omega_j)$



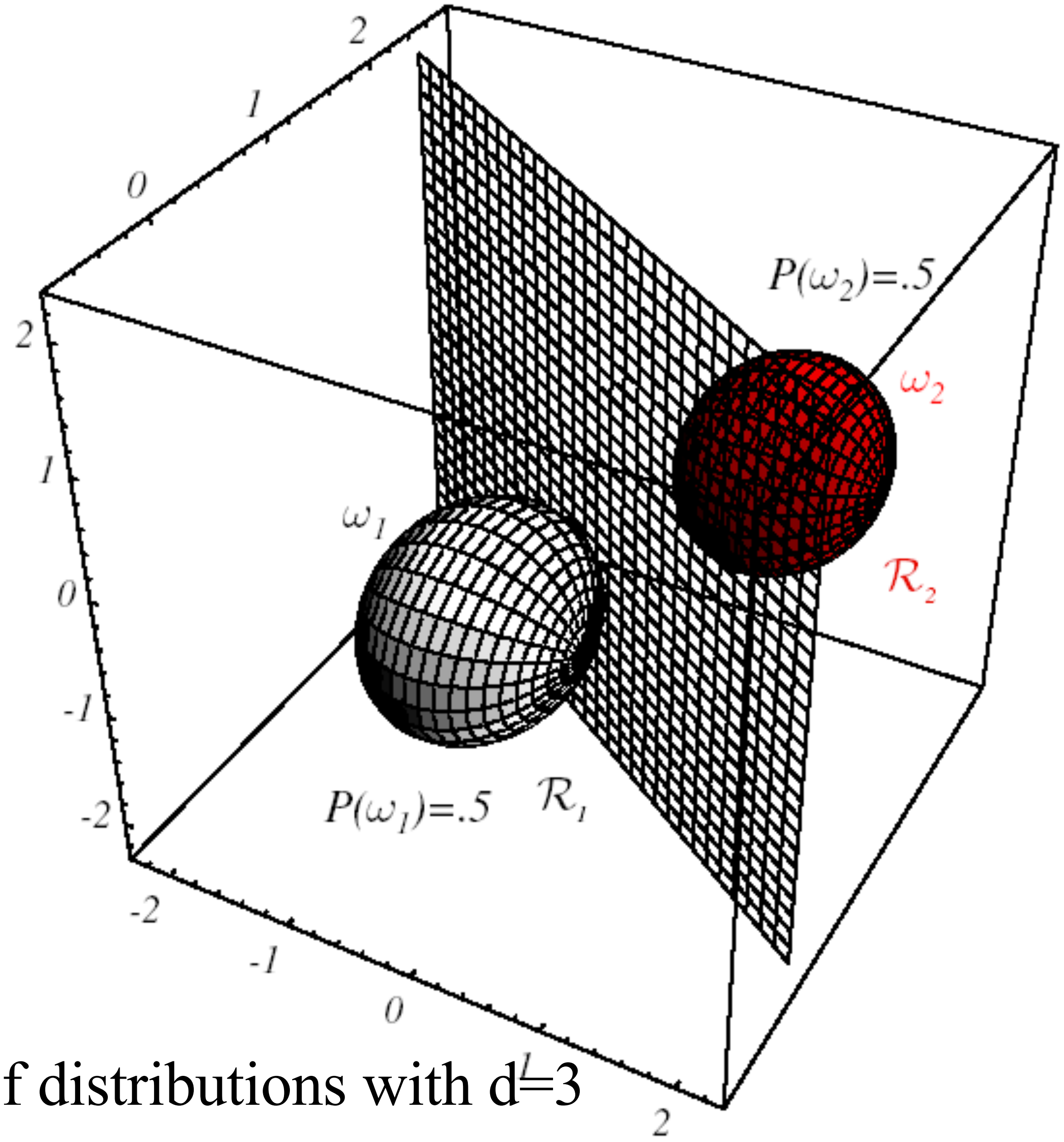
•Example of one-dimensional distributions and relative separation surfaces for the case $\Sigma_i = \sigma^2 I$. In d> 1 dimensions distributions are "spherical", as shown in the next slides.

Example for the case $\Sigma_i=\sigma^2 I$, $P(\omega_i) = P(\omega_j)$



- Example of distributions with d = 2

Example for the case $\Sigma_i=\sigma^2 I$, $P(\omega_i) = P(\omega_j)$

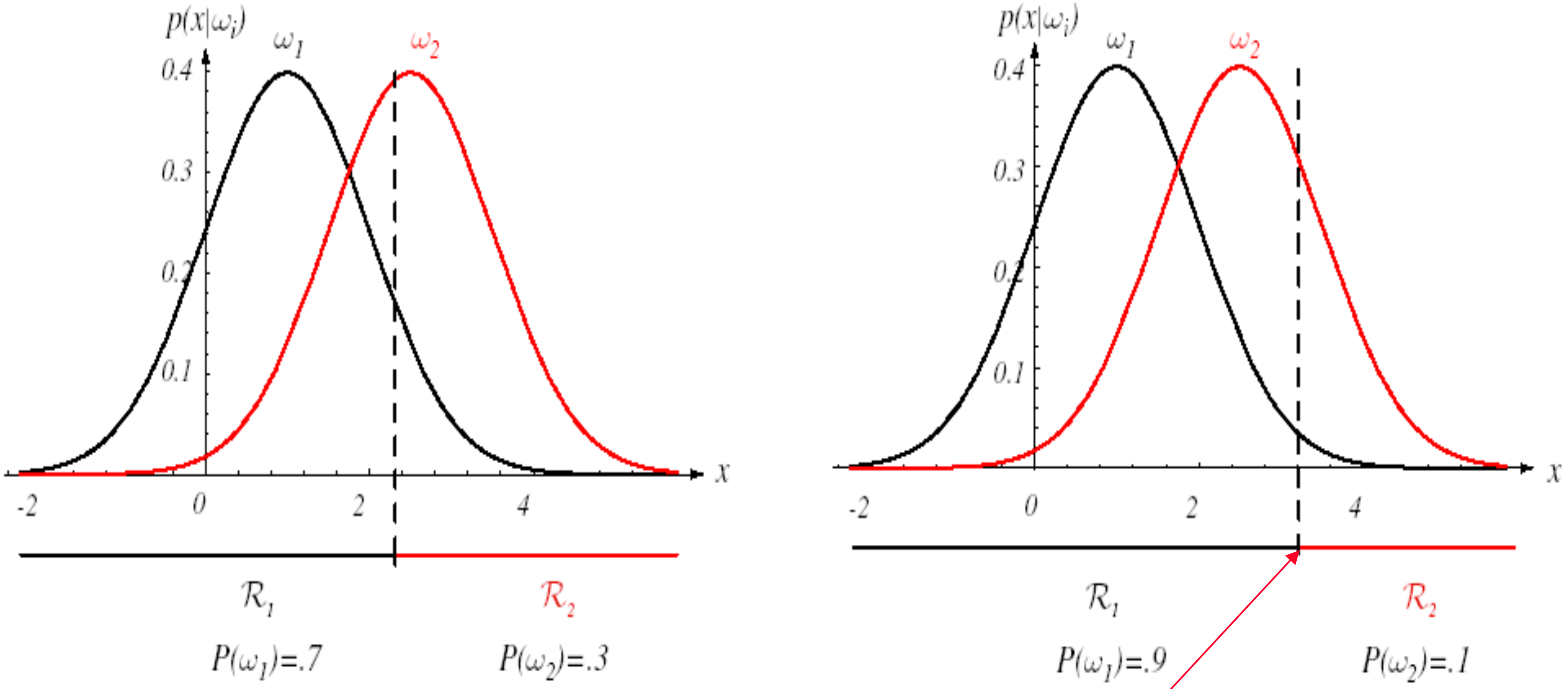


- Example of distributions with d=3

Case $\Sigma_i=\sigma^2 I$ with different a priori probabilities

- As shown in the following slides, changing the prior probabilities changes decision surfaces.
- For prior probabilities quite different, decision boundaries do not lie in between the means of the distributions
- In the next slides, examples for one, two and three dimensions are shown

Example $\Sigma_i=\sigma^2 I$ with different *a priori* probabilities

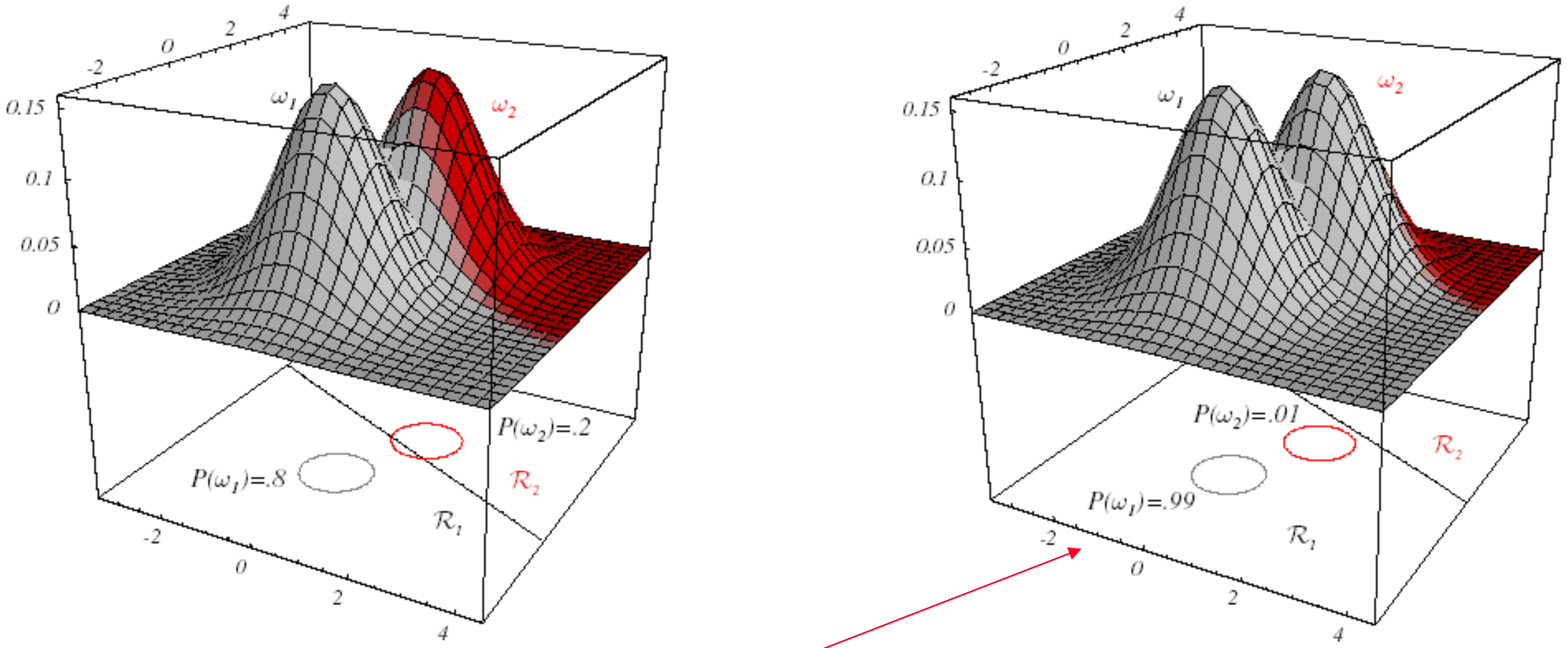


- If the prior probabilities are different the boundary point x_0 "moves away" as is intuitive, from the mean of the most probable class

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

The smaller the variance (compared to the distance between the averages), the more the influence of "priors" decreases

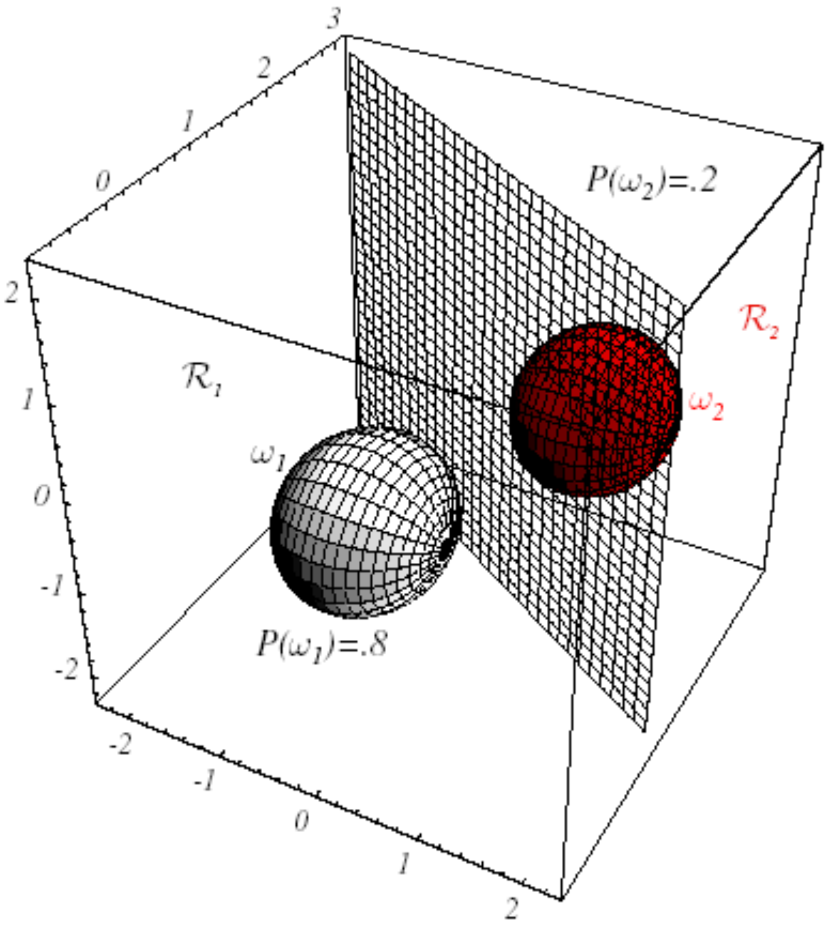
Example $\Sigma_i=\sigma^2 I$ with different *a priori* probabilities

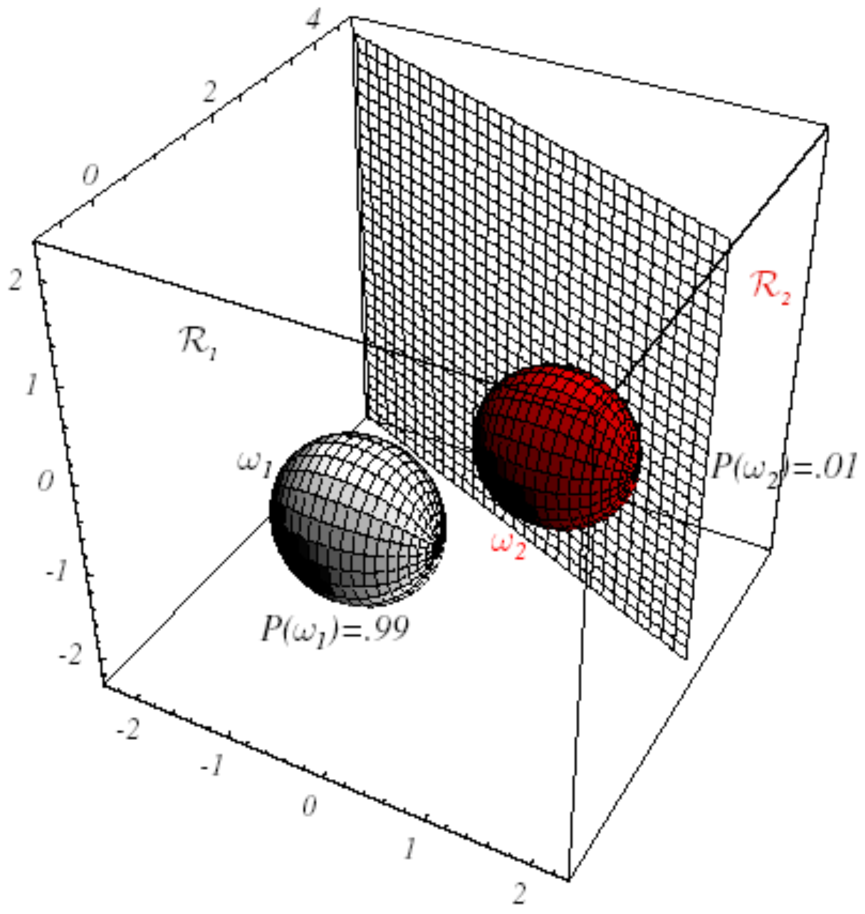


Extreme case where the big difference between $P(\omega_i)$ brings me to choose almost always for ω_1

It is clear that this is a problem when I have to recognize classes “rare” (a rare disease, intrusive traffic, “spamming”)

Example, 3D, $\Sigma_i=\sigma^2 I$, with different *a priori* probabilities





Gaussian model: case $\Sigma_i = \Sigma$

- In this case, the covariance matrices are equal (but arbitrary) for all classes.
- The pattern form hyper-ellipsoidal “clusters” of identical size and shape, centred in μ_i
- Deleting from the discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

all the terms that do not depend on *i*, we can write:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

Gaussian model with $\Sigma_i = \Sigma$ and $P(\omega_i) = P(\omega_j)$

- Special case: $P(\omega_i) = P(\omega_j)$ for each class.

- In this case, the discriminant function becomes:

$$g_i(\mathbf{x}) = -\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}_{\text{distanza di Mahalanobis}}$$

- The decision rule will be:
  - Given **x**, we measure the Mahalanobis distance between **x** and any μ_i, and assign **x** to the class to minimum distance

As in the previous case with Σ diagonal, expanding and eliminating the terms independent on "*i*" we obtain the *linear* function:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

with:

$$w_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i; \quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

Case $\Sigma_i = \Sigma$: Decision Surfaces

- The separation surfaces between adjacent regions R_i and R_j are hyperplanes of equation

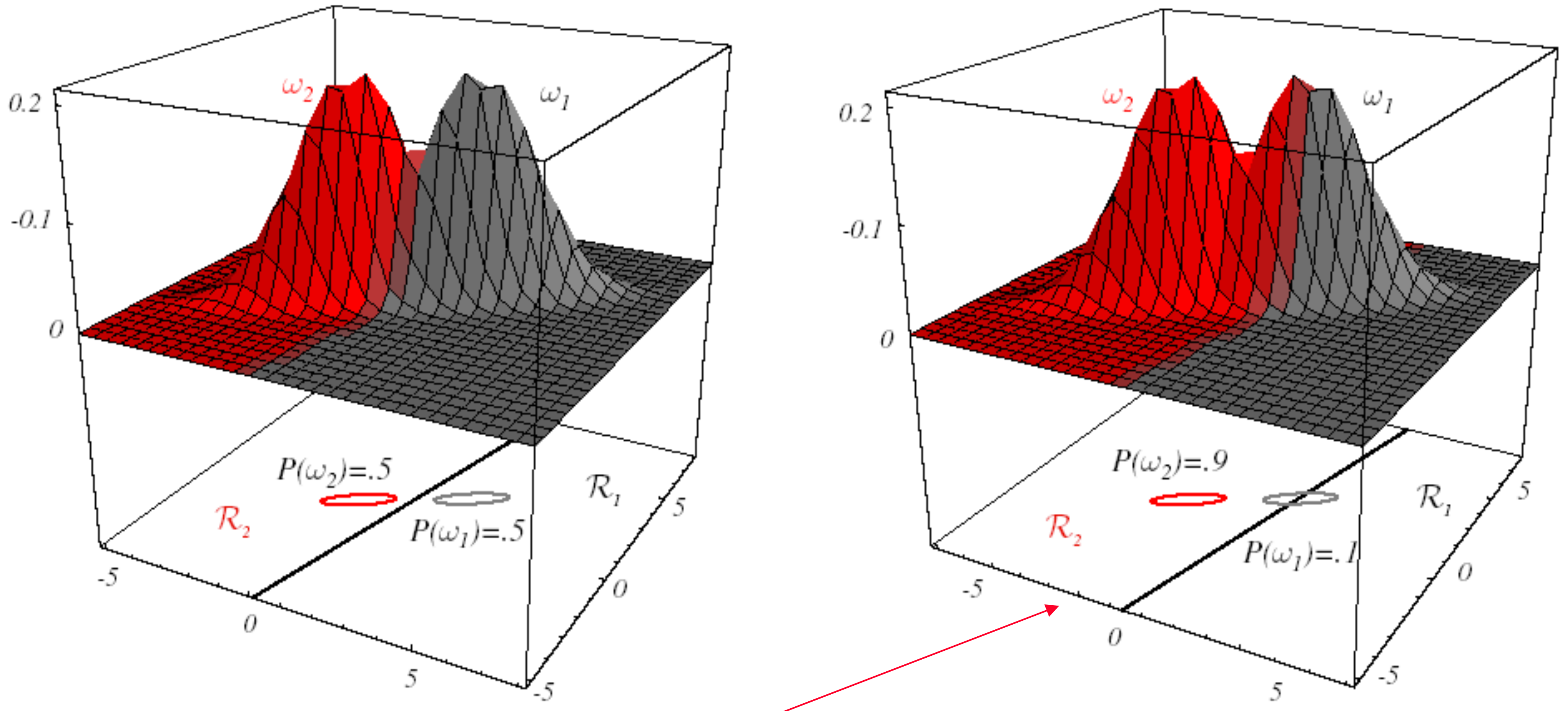
$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

dove

$$\begin{cases} \mathbf{w}=\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right) \\ \mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \end{cases}$$

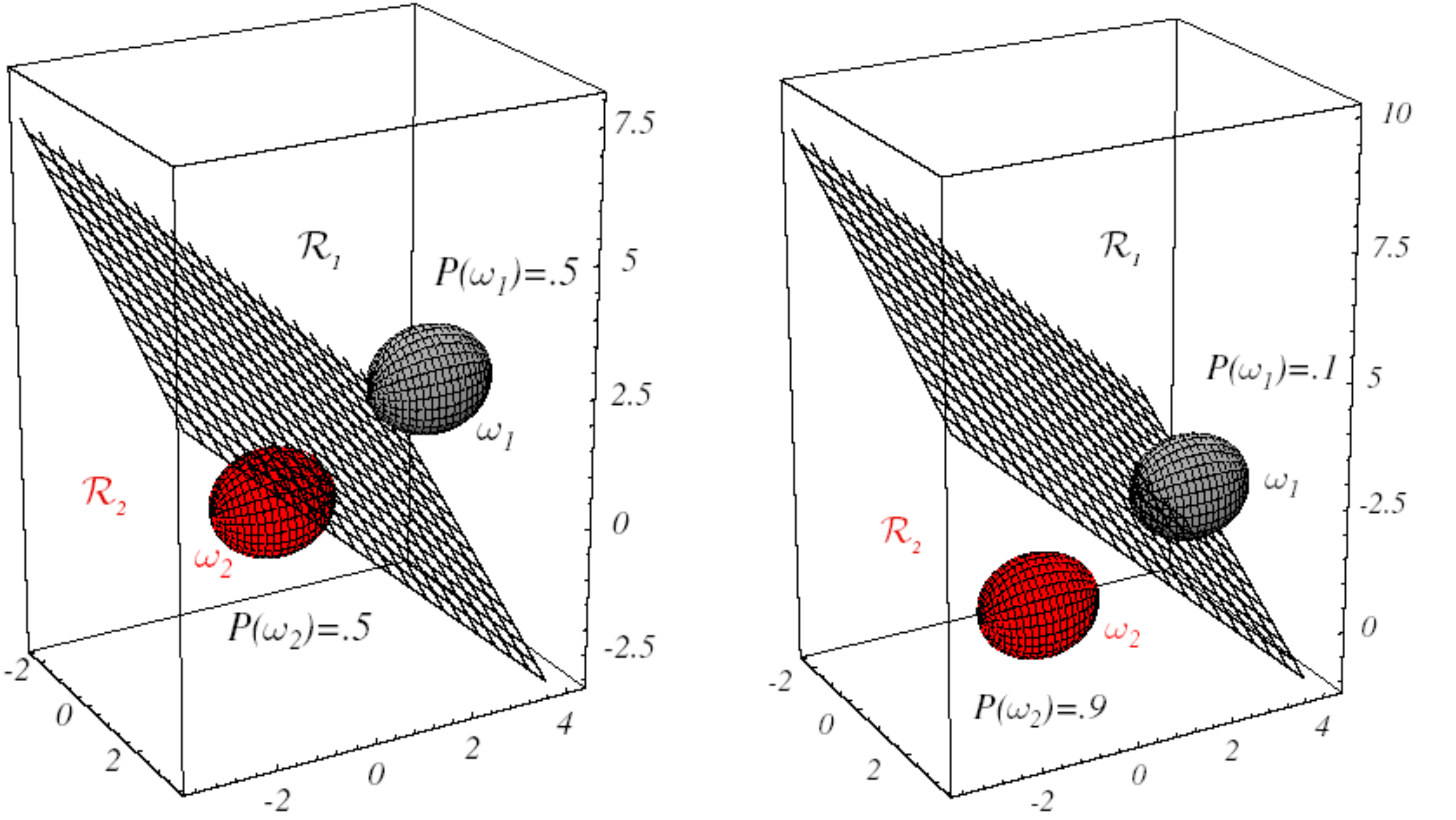
- Since **w** is not (in general) along the direction of (μ_i-μ_j), the hyper-plane is not orthogonal to the line joining the two means.
- However, the hyper-plane intersects this line in $\mathbf{x}_0$; the position of $\mathbf{x}_0$ depends on the *a priori* probability

Example case $\Sigma_i = \Sigma$



- Examples of decision regions for equal normal distributions with very different prior probabilities

Example 3D case $\Sigma_i = \Sigma$



The decision hyperplane may not be orthogonal to the "line" joining the means of classes

Estimation of the covariance matrix if $\Sigma_i = \Sigma$

If the covariance matrix is the same for all classes, it can be shown that its maximum likelihood estimate is:

$$S_w = \sum_{i=1}^{C} \frac{n_i}{n} \hat{\Sigma}_i$$

pooled within-group sample covariance matrix

Where Σ_i is estimated with the samples belonging to class ω_i.

The unbiased estimate of the matrix S_w (for “c” classes) is:

$$\frac{n}{n-c} S_w$$

Gaussian Model: arbitrary Σ_i

- In this case, the only term that we can drop from the discriminant function is (d/2)ln(2π), obtaining:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

- $g_i(\mathbf{x})$ is a *quadratic function* that can be written as

$$g_i(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i'\,\boldsymbol{x} + w_{i0}$$

where

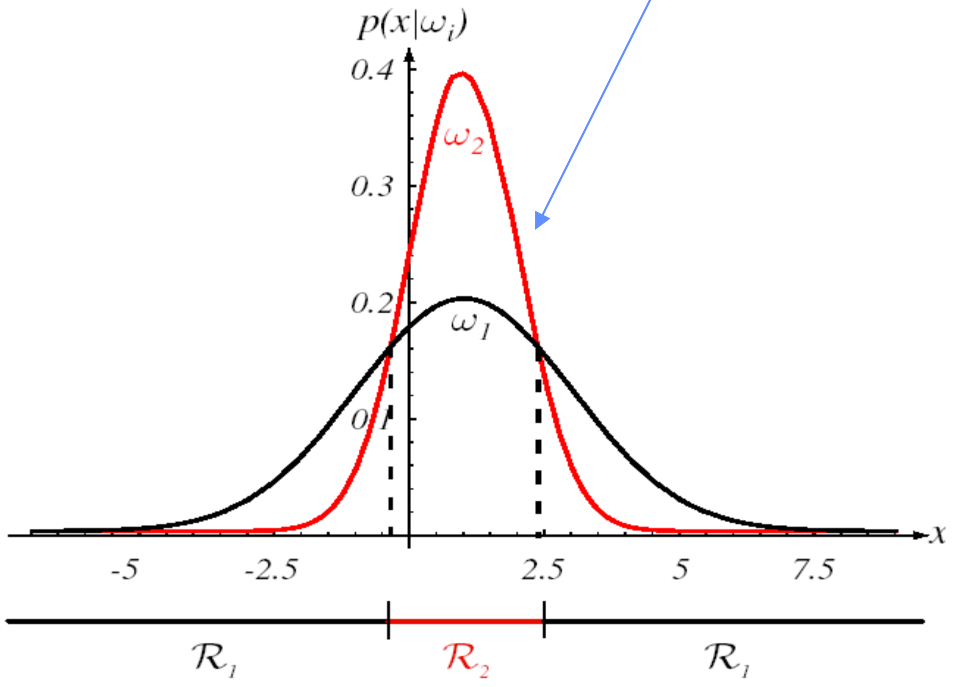
$$\boldsymbol{W}_i = -\frac{1}{2}\,\boldsymbol{\Sigma}_i^{-1}; w_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

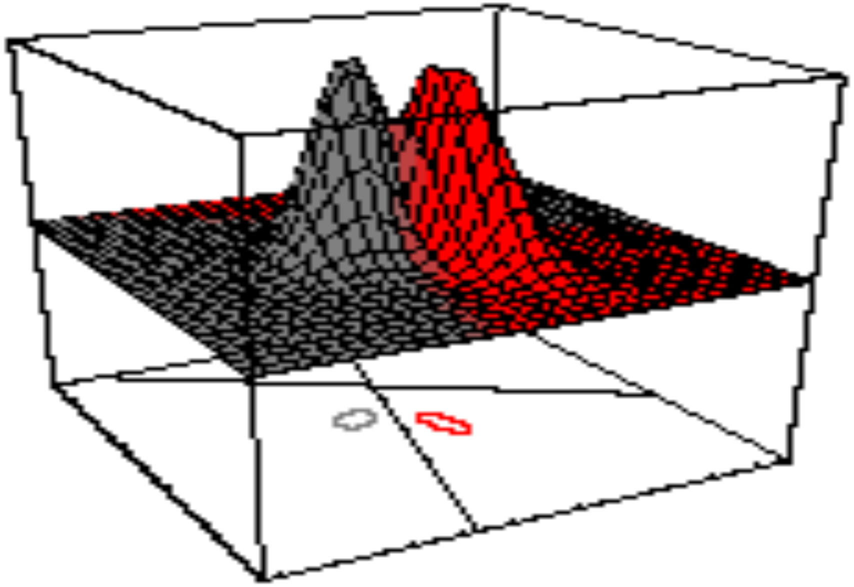
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Gaussian Model: arbitrary Σ_i

Two-class problem:

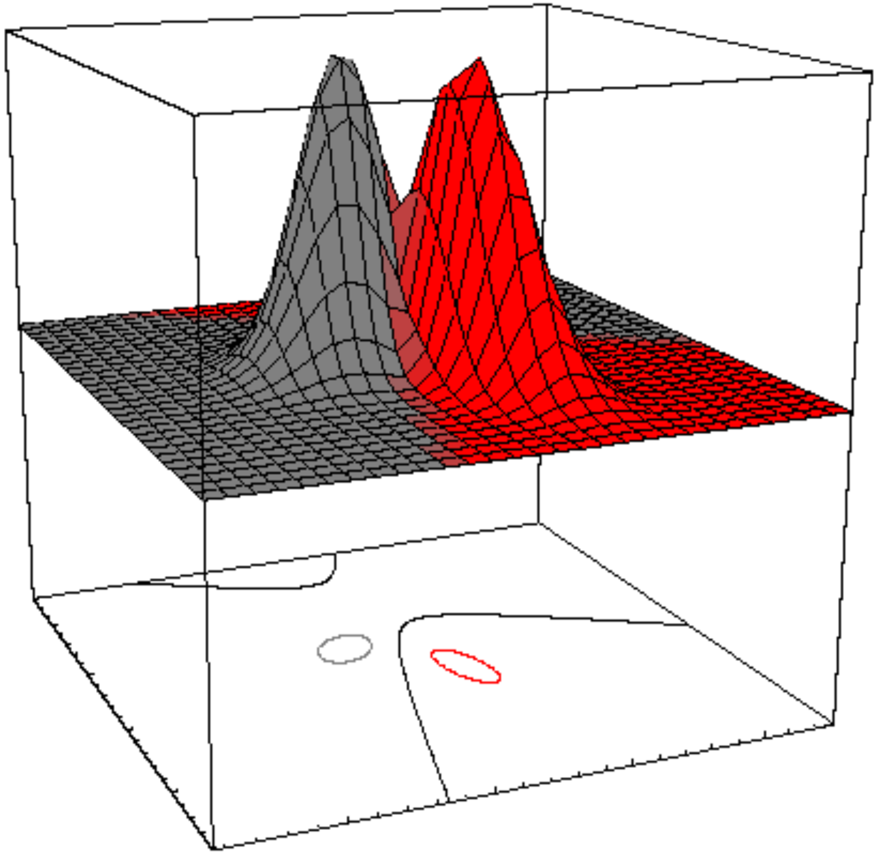
- The decision surfaces are hyperquadrics, and can have any of the general forms. Some examples are given in the following slides.
- Even in one dimension, for an arbitrary covariance matrix the decision regions could not be simply connected.

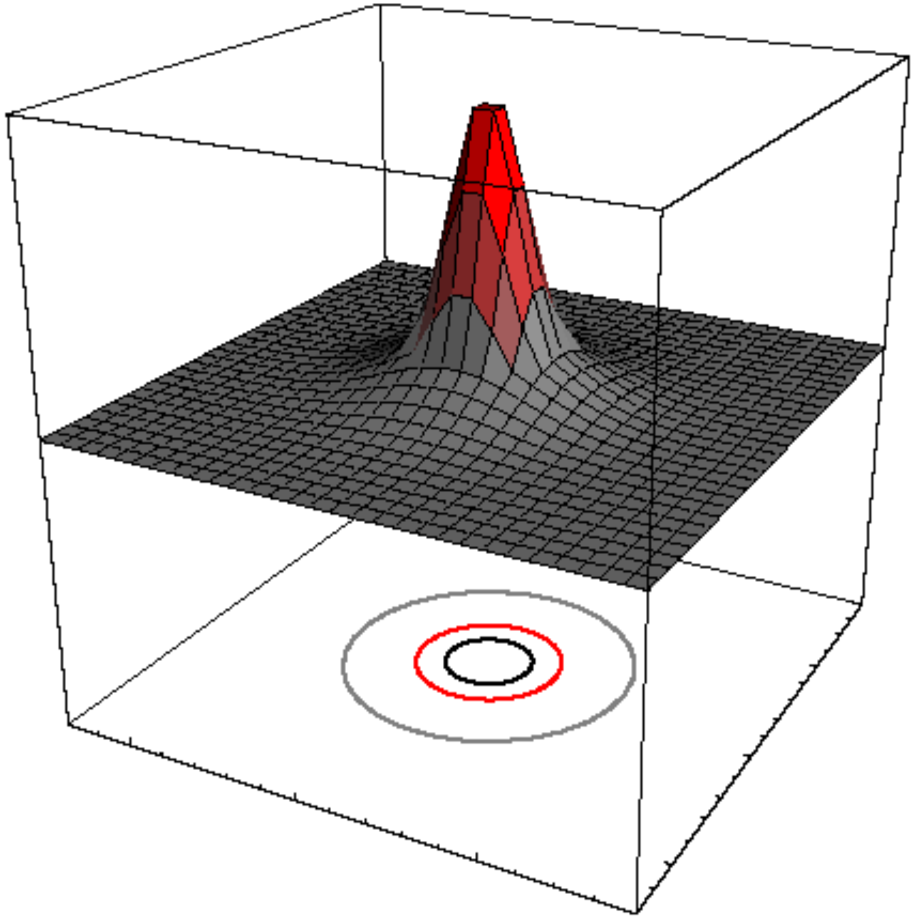




•In problems with *c* classes, we need to keep clear which two of the total classes are responsible for any sub-region of the space.

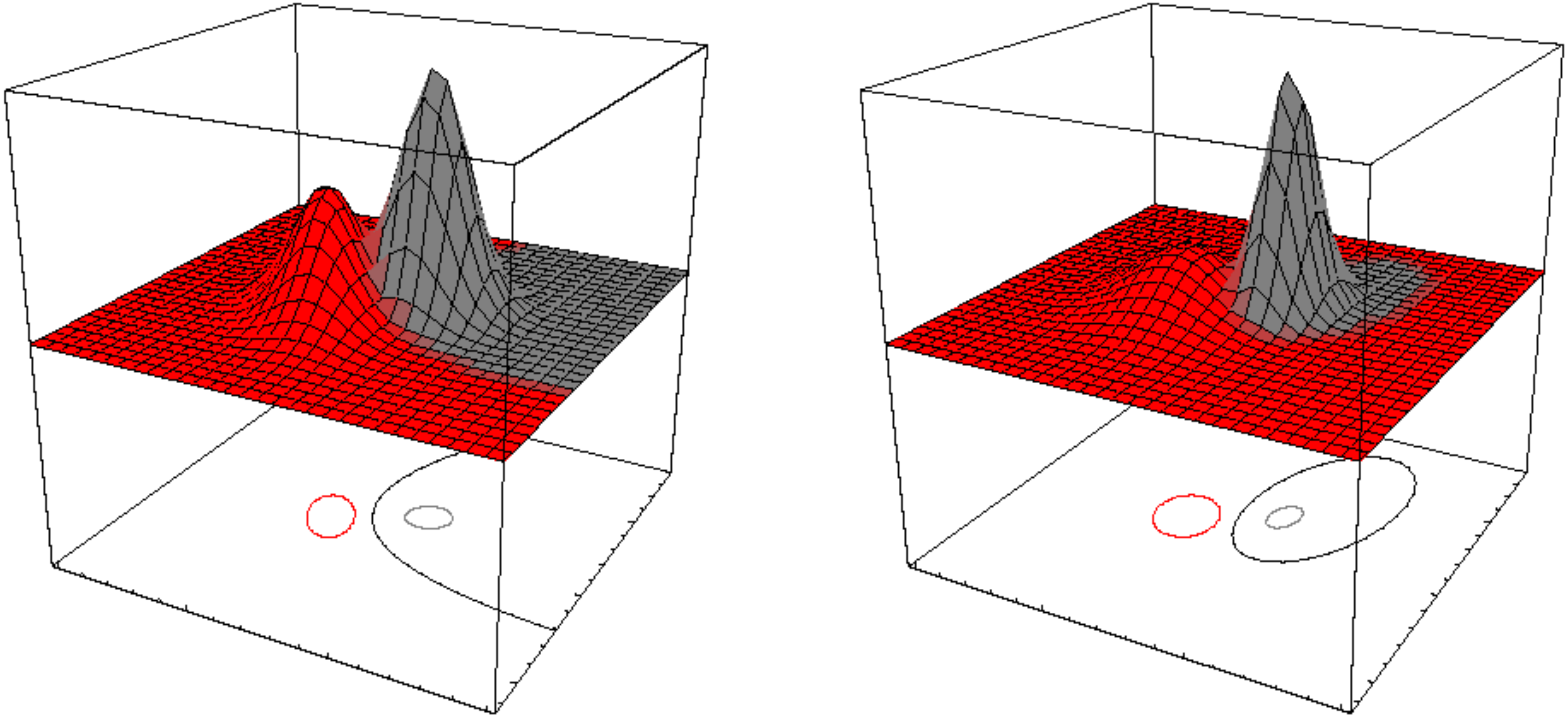
Arbitrary Σ_i : examples





Arbitrary distributions and hyperquadrics decision surfaces

Arbitrary Σ_i : examples



Arbitrary distributions and hyperquadrics decision surfaces

Design of a classifier based on the Gaussian model

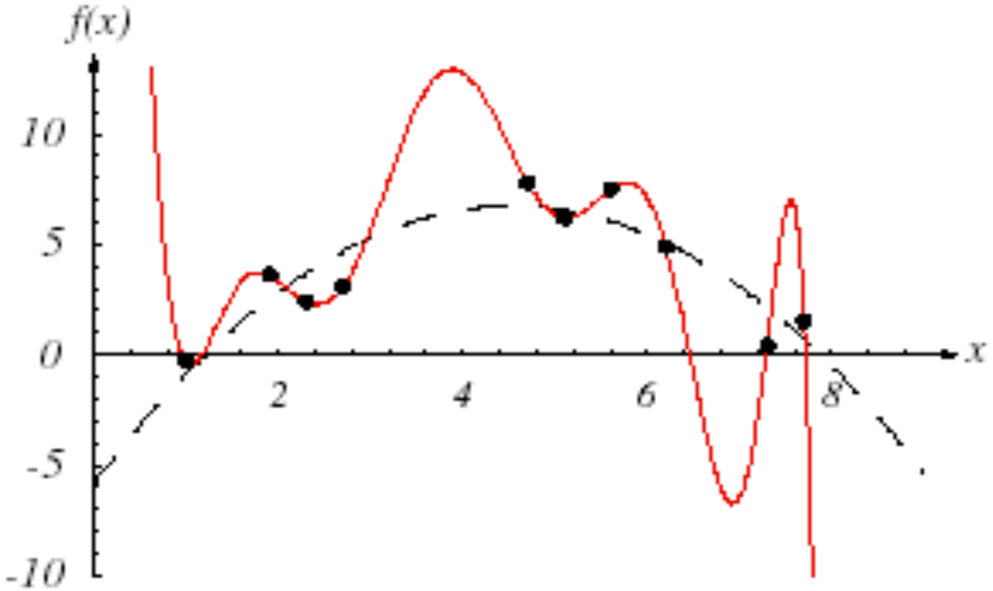
- We should know the kind of covariance matrix Σ_ι we must use to model the distribution of the data

- Usually we do not have such a knowledge

- *One could think that we could use the most general form of* Σ_ι *(the arbitrary case)*

- However, computing the arbitrary Σ_ι requires to estimate d(d+1)/2 parameters, making the estimation unfeasible as the number of features grows
- In the general case of a Gaussian classifier, we need to compute $\boldsymbol{\Sigma}_i^{-1}$
- The inverse matrix is definite only if the number of patterns $n>d$
- For a good estimate of Σ_ι ,we need $n>>d$
- It is hard to say how much patterns n we need, specially in the case of noisy data
- With n “small” and noisy data → “overfitting”

Parameter estimation from noisy data: “Overfitting”



- General problem: a complex model (Σ arbitrary) is more sensible to noise, as it has more parameters

➢In real applications (with noisy data) a classifier based on $\Sigma=\sigma^2 I$ may work better than a classifier with an arbitrary Σ even if we know that the latter model is the correct one

- Analogous problem: curve fitting (regression): the best model for the f(x) plotted on the left is a parabola. But the samples are few and noisy

- Using a more general model (tenth-degree polynomial) we obtain a model that is very different to a parabola (OVERFITTING)

- Similarly, if the right model is $\Sigma=\sigma^2 I$, then using arbitrary Σ can cause overfitting if we do not have enough data

- Overfitting is a problem even if we know the correct model

Parameter estimation from few data ("small sample size")

➢For a good estimate of Σ_ι ,we need $n>>d$

- We can assume that the accuracy of estimation of arbitrary Σ is proportional to

$$1-\frac{d}{n-d}$$

- Accuracy=1 if $n \to +\infty$
- If $\Sigma=\sigma^2 I$ we can assume that the accuracy of estimation is proportional to $1-\frac{d}{n}$
- The two estimates have a similar accuracy if n>>d
- If n is small, the estimation of $\Sigma=\sigma^2 I$ has a higher accuracy

With "few" data with respect to the number of "features" ("small sample size data sets"), "simple" models ($\Sigma=\sigma^2 I$) are more suitable. If n>>d, we can exploit more complex models (Σ arbitrary).

Other single-component models

Many single-component parametric distributions exist. An important family of distributions is the exponential one. It includes the Gaussian distribution, the Poisson distribution, the Rayleigh distribution, and others.

$$p(x \mid \theta) = \alpha(x) \exp[a(\theta) + b(\theta) c(x)]$$

General form of the exponential distribution

θ is a parameter vector

Examples

Simple exponential distribution

$$p(x \mid \theta) = \begin{cases} \theta e^{-\theta x} & \mathrm{x} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Es. the length of network packets approximately follows an exponential distribution

The multinomial distribution is a generalization of the binomial one

Is a model for classifier with “ discrete ” features (each feature can have one of d values with probability θ_i)

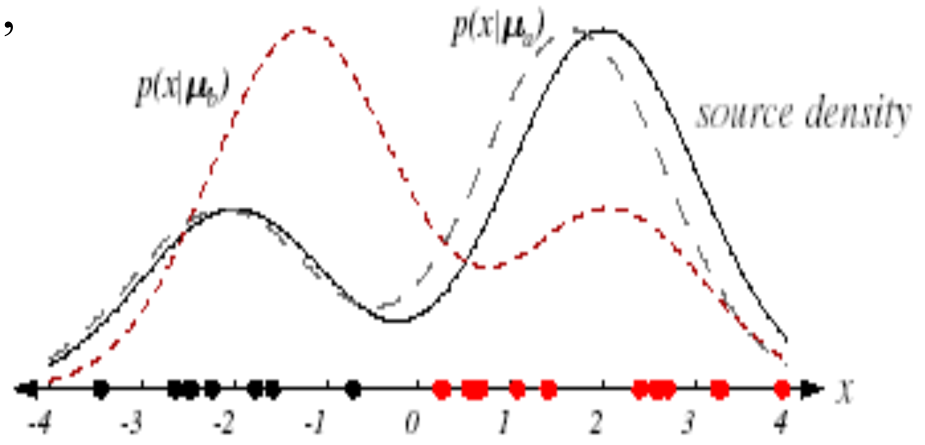
Multinomial Distribution

$$P(x \mid \theta) = \frac{m! \prod_{i=1}^{d} \theta_i^{x_i}}{\prod_{i=1}^{d} x_i !} \qquad \mathrm{x}_\mathrm{i} = 0, 1, \ldots, m \quad 0 < \theta_\mathrm{i} < 1$$

$$\sum_{i=1}^{d} x_i = m \qquad \sum_{i=1}^{d} \theta_i = 1$$

Notes on Multi-component models: Mixtures of distributions

Given *g* distributions $p_1(x), p_2(x), \ldots, p_g(x)$, we define a mixture of distributions as:



"mixing proportions"

$$\sum_{i=0}^{g} p_i p_i(x)$$

with $p_i \geq 0 \quad \sum_{i=0}^{g} p_i = 1$

- a mixture can define more complex models than the source distribution
- For example, a mixture of two Gaussians is a distribution with five parameters, and can represent also bimodal distributions
- The mixture model is good whenever the classes are multimodal, or have multiple "components"
- With mixtures, the number of parameters to be estimated grows. If each $p_i(x)$ has *p* parameters, the mixture with *g* components has [(p+1)g – 1] parameters, that have to be estimated

Notes on use of multi-component models

- The mixture models are useful when patterns of a single class belong to different "groups", and we know the class of the pattern, but not the group within the class.

- Example: text categorization

- A certain text category (e.g. sport) may be a mixture of different specific sub-categories (e.g. soccer, tennis, etc.). Each specific category may need to be modelled in a different way.

- Text categorization is based on "features" that represent the frequency (count of the occurrences) of certain keywords. Different sub-categories of the class "Sport" are likely to show different "co-occurrences" of the keywords. It is easy to see that such problem can be better modelled with a multiple-component model.

- If we consider an unique, general category "Sport", we do not know to what specific sub-category (e.g. Soccer, Tennis, etc.) the pattern belongs. **This can make the problem of parameter estimation harder**.

References

- Sections 2.6, 3.1, 3.2, 3.3, 3.9, Pattern Classification, R. O. Duda, P. E. Hart, D. G. Stork, John Wiley & Sons, 2000.
- Sections 2.1, 2.2.1, 2.2.2, 2.3, Statistical Pattern Recognition, Andrew Webb, John Wiley & Sons, 2002.